IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE HONORABLE BOARD OF PATENT APPEALS AND INTERFERENCES

In re the Application of

James M. Sweet et al.

Application No.: 10/608,591

Filed: 06/27/2003

Confirmation No. 8445

Examiner: Nathan Hillery

Docket No.: A2555Q1-US-NP

For: **DETERMINATION OF TABLE OF CONTENT LINKS FOR A HYPERLINKED DOCUMENT**

BRIEF ON APPEAL

Appeal from Group 2176

Christopher D. Wait
XEROX CORPORATION
Xerox Square - 20A
Rochester, NY 14644
Telephone: (585) 423-6918
Attorneys for Appellants

TABLE OF CONTENTS

I.   <u>REAL PARTY IN INTEREST</u>

The real party in interest for this appeal and the present application is Xerox

Corporation, by way of an Assignment recorded in the U.S. Patent and Trademark Office at

Reel 14257, Frame 925-927 and Reel 014557, Frame 0676-0677.

II. STATEMENT OF RELATED APPEALS AND INTERFERENCES

Following are identified any prior or pending appeals, interferences or judicial proceedings, known to Appellant, Appellant's representative, or the Assignee, that may be related to, or which will directly affect or be directly affected by or have a bearing upon the Board's decision in the pending appeal:

Appeal Brief filed in copending Application No. 10/608,590

Appeal Brief to be filed in copending Application No. 10/608,587

There are no further prior or pending appeals, interferences or judicial proceedings, known to Appellant, Appellant's representative, or the Assignee, that may be related to, or which will directly affect or be directly affected by or have a bearing upon the Board's decision in the pending appeal.

III.   <u>STATUS OF CLAIMS</u>

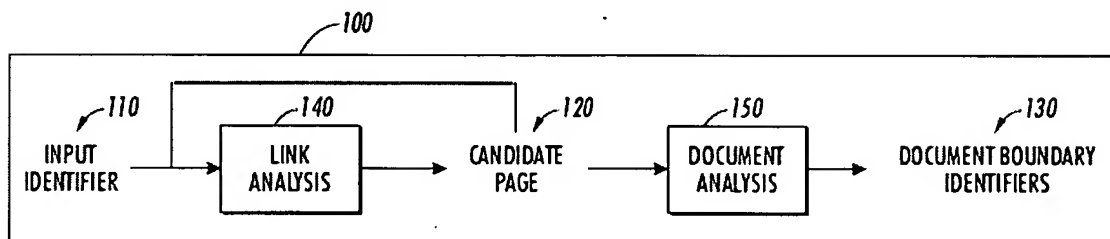Claims 1, 2, 4, 6, 7, 9, 11, 12, and 14 are on appeal.

Claims 1-15 are rejected.

IV.   STATUS OF AMENDMENTS

No Amendment After Final Rejection has been filed.

V.   SUMMARY OF CLAIMED SUBJECT MATTER

The subject matter of independent claim 1 is directed to a methodology for assembling a document from content spanning multiple web-pages by employing two cooperative processes. Given a starting location 110, one process analyzes a single page at a time to find candidate links 140. The links are recursively followed and those pages are analyzed. A detailed set of heuristics is used to determine what is or is not a candidate link. The links are examined for link clusters and a table of contents if found is identified. The candidate pages 120 are then fed to a document-level analyzer 150. This process compares the attributes of one page against the others and looks for a document-like structure. Using another detailed set of heuristics, the document-level analyzer 150 determines if the page should be included in the document. (see Abstract, page 20 of the specification as filed, and Figure 1) [in support of claim 1]
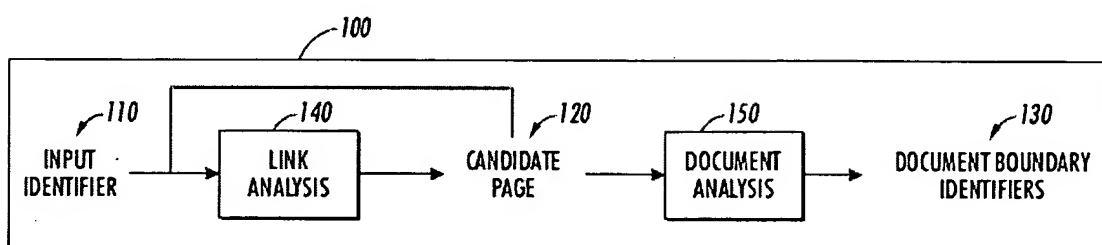

FIG. 1

The subject matter of independent claim 6 is directed to a methodology for assembling a document from content spanning multiple web-pages by employing two cooperative processes. Given a starting location 110, one process analyzes a single page at a time to find candidate links 140. The links are recursively followed and those pages are analyzed. A detailed set of heuristics is used to determine what is or is not a

candidate link. The links are examined for link clusters and a table of contents if found is identified. The candidate pages 120 are then fed to a document-level analyzer 150. This process compares the attributes of one page against the others and looks for a document-like structure. Using another detailed set of heuristics, the document-level analyzer 150 determines if the page should be included in the document. (see Abstract, page 20 of the specification as filed, and Figure 1) [in support of claim 6]

FIG. 1

In more particular support of the subject matter of independent claim 6, the page-level link analysis 140 is described in greater detail in Figure 2. During page-level link analysis 140, the document detection system attempts to identify links that may potentially lead to other pages within the same document. It is assumed that a well-authored multi-page document will always include progression links (links that provide some well-defined progression through the document, often indicated by the presence of some well-known contextual clue, such as a graphic or text "next" or "previous" indicator) and/or table of contents links (clusters of links providing a path to every page or some logical subset of pages in the document) that indicate the structure of the document. These are the two categories of intra-document links that the link analysis process 140 seeks to identify. (see page 7, lines 10-20 of the specification as filed, and Figure 2) [in support of claim 6]
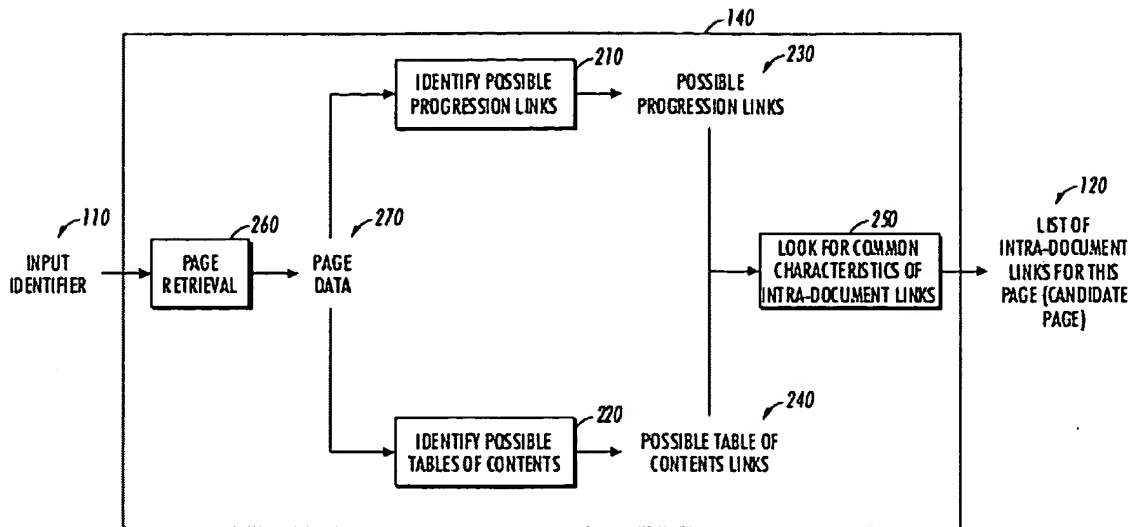
FIG. 2

In further support of the subject matter of independent claim 6, the link analysis process begins with the retrieval of the actual page 270 for analysis from the page identifier 110. This is done as will be well understood by those skilled in the art, by the page retrieval process 260. The retrieved page 270 is then used as input to both the progression-link identification module 210 and the link-cluster identification module 220. In the progression-link identification module 210, possible progression links 230 are identified primarily by means of a progression indicator, which is a textual or graphical clue that suggests the nature of the progression link. Link-cluster identification module 220 examines the page data 270 to identify link clusters and thereby possible table of content type links 240. The possible progression links 230 and possible table of content links 240 are passed to module 250 for a final examination to weed out links which have properties that are not characteristic of typical intra-document links, e.g. they point to a different web server. The final result is then a list of intra-document links 120 for the candidate page 270. (see page 7, lines 22-35 of the specification as filed, and Figure 2)

Figure 2, module 220 examines the page data 270 to identify link clusters. It is assumed that in a well-authored hypertext page, table of contents links will appear in clusters, thereby indicating to the user that all of these links are part of a single cohesive construct. Given this assumption, the first step in locating a table of contents is to locate all of the link clusters in a particular page.

The Identification of link clusters is based on three criteria:

1) Proximity: The links in a cluster should be close together. The same heuristic as applied to identification of the most proximal link for a progression indicator can be used here to identify groups of links that have a low perceived distance.

2) Similarity: The links in a cluster should look like each other, i.e. they will usually all be of the same font, type size, and color.
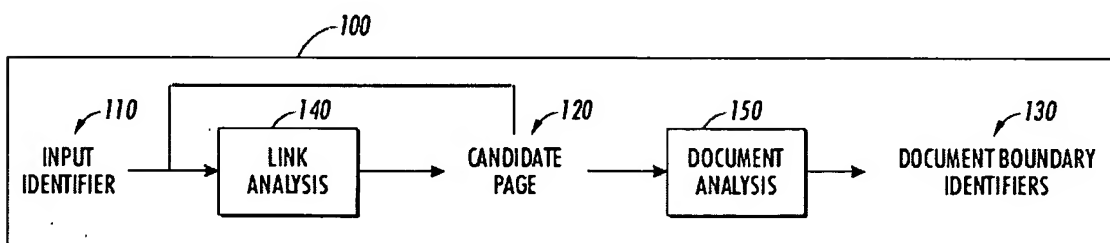
3) Regularity: If there is intervening content between the links, or if the links are dissimilar, these lapses in Proximity and Similarity should form some sort of consistent pattern. One example is a table of contents where each link has a chapter description below it (Proximity is low, but the pattern of intervening content is highly consistent). Another example is a table of links where the color of the text alternates in each column in order to make it more readable (Similarity is low, but the changes in appearance form a simple pattern).

Regularity is measured by performing pattern matching on the intervening content and document structure tags between pairs of nearby links. The other two criteria are easily measured by simple heuristics.

Once all link clusters in a web page have been identified, the task remains of distinguishing which clusters represent tables of contents and which represent other constructs, such as navigation bars or bibliographies. The primary determining criteria

8

for this is the similarity between the link targets of the links in the cluster, i.e. collocation on the same server, residence in the same directory or nearby area of the directory hierarchy, and similarity in filename. (see page 10, lines 4-33 of the specification as filed) [in support of claim 6]

The subject matter of independent claim 11, is directed to a methodology for assembling a document from content spanning multiple web-pages by employing two cooperative processes. Given a starting location 110, one process analyzes a single page at a time to find candidate links 140. The links are recursively followed and those pages are analyzed. A detailed set of heuristics is used to determine what is or is not a candidate link. The links are examined for link clusters and a table of contents if found is identified. The candidate pages 120 are then fed to a document-level analyzer 150. This process compares the attributes of one page against the others and looks for a document-like structure. Using another detailed set of heuristics, the document-level analyzer 150 determines if the page should be included in the document. (see Abstract, page 20 of the specification as filed, and Figure 1) [in support of claim 11]

FIG. 1

In more particular support of the subject matter of independent claim 11, the page-level link analysis 140 is described in greater detail in Figure 2. During page-level

link analysis 140, the document detection system attempts to identify links that may potentially lead to other pages within the same document. It is assumed that a well-authored multi-page document will always include progression links (links that provide some well-defined progression through the document, often indicated by the presence of some well-known contextual clue, such as a graphic or text "next" or "previous" indicator) and/or table of contents links (clusters of links providing a path to every page or some logical subset of pages in the document) that indicate the structure of the document. These are the two categories of intra-document links that the link analysis process 140 seeks to identify. (see page 7, lines 10-20 of the specification as filed, and Figure 2) [in support of claim 11]
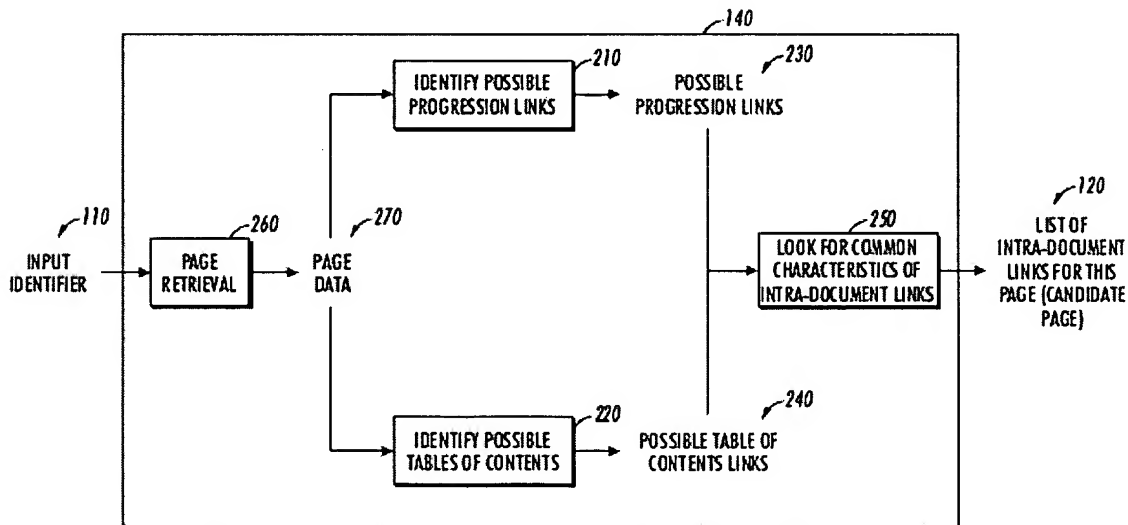

FIG. 2

In further support of the subject matter of independent claim 11, the link analysis process begins with the retrieval of the actual page 270 for analysis from the page identifier 110. This is done as will be well understood by those skilled in the art, by the page retrieval process 260. The retrieved page 270 is then used as input to both the

progression-link identification module 210 and the link-cluster identification module 220. In the progression-link identification module 210, possible progression links 230 are identified primarily by means of a progression indicator, which is a textual or graphical clue that suggests the nature of the progression link. Link-cluster identification module 220 examines the page data 270 to identify link clusters and thereby possible table of content type links 240. The possible progression links 230 and possible table of content links 240 are passed to module 250 for a final examination to weed out links which have properties that are not characteristic of typical intra-document links, e.g. they point to a different web server. The final result is then a list of intra-document links 120 for the candidate page 270. (see page 7, lines 22-35 of the specification as filed, and Figure 2)

Figure 2, module 220 examines the page data 270 to identify link clusters. It is assumed that in a well-authored hypertext page, table of contents links will appear in clusters, thereby indicating to the user that all of these links are part of a single cohesive construct. Given this assumption, the first step in locating a table of contents is to locate all of the link clusters in a particular page.

The Identification of link clusters is based on three criteria:

1) Proximity: The links in a cluster should be close together. The same heuristic as applied to identification of the most proximal link for a progression indicator can be used here to identify groups of links that have a low perceived distance.

2) Similarity: The links in a cluster should look like each other, i.e. they will usually all be of the same font, type size, and color.

3) Regularity: If there is intervening content between the links, or if the links are dissimilar, these lapses in Proximity and Similarity should form some sort of consistent pattern. One example is a table of contents where each link has a chapter

description below it (Proximity is low, but the pattern of intervening content is highly consistent). Another example is a table of links where the color of the text alternates in each column in order to make it more readable (Similarity is low, but the changes in appearance form a simple pattern).

Regularity is measured by performing pattern matching on the intervening content and document structure tags between pairs of nearby links. The other two criteria are easily measured by simple heuristics.

Once all link clusters in a web page have been identified, the task remains of distinguishing which clusters represent tables of contents and which represent other constructs, such as navigation bars or bibliographies. The primary determining criteria for this is the similarity between the link targets of the links in the cluster, i.e. collocation on the same server, residence in the same directory or nearby area of the directory hierarchy, and similarity in filename. (see page 10, lines 4-33 of the specification as filed) [in support of claim 11]

VI.   GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

The following grounds of rejection are presented for review:

Claims 1, 2, 4, 6, 7, 9, 11, 12, and 14 are rejected under 35 U.S.C. §103(a) as being unpatentable over U.S. Patent No. 6,112,203 to Bharat et al. (hereinafter Bharat) in further view of U.S. Patent No. 5,924,104 to Earl (hereinafter Earl).

Claims 3, 5, 8, 10, 13, and 15 are rejected under 35 U.S.C. §103(a) as being unpatentable over Bharat and Earl, in further view of U.S. Patent No. 6,633,868 to Min et al. (hereinafter Min).

VII. ARGUMENT

A. Claims 1, 2, 4, 6, 7, 9, 11, 12, and 14 Would Not Have Been Obvious Over Bharat in View of Earl

Claims 1, 2, 4, 6, 7, 9, 11, 12, and 14 are rejected under 35 U.S.C. §103(a) as being unpatentable over Bharat in further view of Earl.

Problematically, neither Bharat or Earl, alone or in combination, teach or suggest the Applicants' invention. Claim elements are missing from the Bharat and Earl references. Indeed the references teach away from the Applicants' claimed invention. A Prima facie case for Obviousness has thus not been made out. Further, no finding has been provided directed to: a identifiable reason that would have prompted a person of ordinary skill in the relevant field to combine the elements in the way that the Applicants' claimed new invention does. Thus, the Applicant is faced with the conundrum of positively proving a negative. That is in other words: proving that something which is not there, is not there.

Bharat teaches that in a computerized method, a set of documents is ranked according to their content and their connectivity by using topic distillation. The documents include links that connect the documents to each other, either directly, or indirectly. A graph is constructed in a memory of a computer system. In the graph, nodes represent the documents, and directed edges represent the links. Based on the number of links connecting the various nodes, a subset of documents is selected to form a topic. A second subset of the documents is chosen based on the number of directed edges connecting the nodes. Nodes in

14

the second subset are compared with the topic to determine similarity to the topic, and a relevance weight is correspondingly assigned to each node. Nodes in the second subset having a relevance weight less than a predetermined threshold are pruned from the graph. The documents represented by the remaining nodes in the graph are ranked by connectivity based ranking scheme.

It is essential to the understanding Bharat that *Bharat is directed to a search engine* and as such is sorting through pages already identified by a simple word string search (please see column 1, lines 14-54 of Bharat). Bharat is concerned with solving the problem of answering a search engine query, and thus with ranking <u>a set of documents</u> to point to in response to that query. The Applicants however, are teaching that having identified where one document page is, how to find and pull together all relevant pages associated with that document into a single coherent document (please see page 5, first paragraph, of the Applicants' specification). That is, a single coherent document representation suitable for printing and downloaded viewing. As such the Applicants teach "to weed out links which have properties that are not characteristic of *intra*-document links" and thus eschew all other documents. Bharat on the other hand, will not link (i.e. Bharat will reject) self referencing pages so as not to unduly influence the search outcome (see column 5, lines 17-20) where Bharat provides:

> "If a link points to a page that is represented by a node in the graph, and both pages are on different servers, then a corresponding edge 213 is added to the graph 211. ***Nodes representing pages on the same server are <u>not</u> linked. This prevents a single Web site with many self-***

referencing pages to unduly influence the outcome. This completes

the n-graph 211."

Thus Bharat is interested in only *inter*-document links for the sake of ranking

links. Bharat does not assemble a single coherent document but a link list of

search results responsive to a word query. Thus Bharat does NOT examine "the

collective set of identified candidate document pages to weed out links which

have properties that are not characteristic of *intra*-document links". Thus a claim

element is missing.

Indeed, Bharat teaches away from the Applicants' invention. The

Applicants teach to embrace that which Bharat discards. The cited text from the

Applicants' specification page 7, lines 22-35 follows:

"The link analysis process begins with the retrieval of the actual

page 270 for analysis from the page identifier 110. This is done as will be

well understood by those skilled in the art, by the page retrieval process

260. The retrieved page 270 is then used as input to both the

progression-link identification module 210 and the link-cluster identification

module 220. In the progression-link identification module 210, possible

progression links 230 are identified primarily by means of a progression

indicator, which is a textual or graphical clue that suggests the nature of

the progression link. Link-cluster identification module 220 examines the

page data 270 to identify link clusters and thereby possible table of

content type links 240. The possible progression links 230 and possible

table of content links 240 are passed to module 250 for a final examination

to weed out links which have properties that are not characteristic of

typical intra-document links, e.g. they point to a different web server. The final result is then a list of intra-document links 120 for the candidate page 270."

To paraphrase, the possible links are passed by the Applicants, to weed out those links which are not characteristic of typical intra-document links. An example of the links which are weeded out are those which point to a different web server. (please see also page 10, lines 32-34, of the Application Specification as filed). These links are not likely to be part of the document the Applicants are trying to assemble. Bharat does just the opposite, as noted above, so as to not unduly influence the outcome of the results to the user query. Please also see the attached §132 Declarations, particularly in the Sweet Declaration, item numbers 7-11, and in the Harrington Declaration, item number 7.

Earl fails to provide what Bharat lacks, nor does it provide any teaching relating to the Applicants' claimed invention. Earl provides link lists like Bharat but provides different presentation styles for the links to a user depending on whether they are intra-document or inter-document. Actually what Earl defines as intra-document is what the Applicants would call intra-page, i.e. a link pointing to a location somewhere further down the same page. And thus what Earl calls inter-document is really inter-page. The teaching found in Earl is simply about providing some indicia to the viewer as to whether a hyper link will take the user elsewhere down the present page or to an entirely different page.

Please see the attached §132 Declarations for what one skilled in the art would consider to be "intra-document" versus intra-page, particularly in the Sweet

17

Declaration, item numbers 13-15, and in the Harrington Declaration, item number 5. The Applicants are teaching assembling a document, and having identified the current page, have no interest in self referential links to that same page (they already have it) and thus would discard, or weed out, those links which Earl .keeps.

Earl, having made a discrimination between two type of links, keeps all those links, choosing only to display them differently. The Applicants having discriminated between links to find some as not pointing to more of the desired document, discard or weed out or filter out those links. A gardener, weeding out a flower bed and having spotted a weed, does not keep that weed in their flower bed to display differently. But Earl does. Thus Earl does NOT examine "the collective set of identified candidate document pages to weed out links which have properties that are not characteristic of typical intra-document links".

In rebuttal to this previously presented argument above analogizing the terminology "weed out" to gardening, the Examiner has asserted that the Office "is forced to rely on the knowledge of one of ordinary skill in the art". The Applicants must emphatically traverse as to how one skilled in the art would interpret this terminology. Please see the attached §132 Declarations particularly the Sweet Declaration, item number 8 and especially item 9 for what those skilled in the art would consider to be meant by the terminology "weed out". Earl only discriminates, but does not discard, thus Earl does not "weed out".

It must also be pointed out that neither Bharat or Earl concern themselves with the claim element of a table of contents. This is again not a surprising finding as they are both directed to search inquiries rather than the singular

18

document image which the Applicants endeavor to assemble. Finding a table of contents is an important find in tracing out a single hyperlinked document, but of little consequence to a search engine inquiry. Thus yet another of the Applicants' claim elements is absent in the cited art of Bharat and Earl.

Therefore, Earl in turn fails to provide the elements that Bharat also lacks, and the combination of Bharat and Earl thus fails to provide the requirements for a Prima Facie case of obviousness and the rejection is improper.

Further, no finding has been provided by the Examiner directed to some *genuine* identifiable reason that would have prompted a person of ordinary skill in the relevant field to combine the elements in the way that the Applicants' claimed new invention does. The importance of doing so is clearly stated in *KSR*, 550 U.S., 82 USPQ2d at 1395 and 1396. It would appear that the Examiner is using Appellant's disclosure as a recipe for selecting the appropriate portions of the prior art to construct Appellant's claimed invention. A piecemeal reconstruction of prior art patents in light of Appellant's disclosure should not be a basis for a holding of obviousness, especially when claim elements are absent.

VIII.   CONCLUSION

For all of the reasons discussed above, it is respectfully submitted that the

rejections are in error and that claims 1, 2, 4, 6, 7, 9, 11, 12, and 14 are in condition for

allowance.   While claims 3, 5, 8, 10, 13, and 15 have not been separately argued, they

depend from claims deemed allowable, thus they should be allowable as well.   For all of

the above reasons, Appellants respectfully request this Honorable Board to reverse the

rejections of claims 1-15.

Respectfully submitted,

/Christopher D. Wait, Reg. #43230/
Christopher D. Wait
Attorney for Applicant(s)
Registration No. 43,230
Telephone (585) 423-6918

XEROX CORPORATION
Xerox Square – 20A
Rochester, NY 14644

Filed: December 9, 2008

20

CLAIMS APPENDIX

CLAIMS INVOLVED IN THE APPEAL:

1.   (Previously Presented)   An automated identification methodology for identification of table of content links in a given hyperdocument for assembling a document representation by gathering the content of hyperlinked pages pointed to by the identified table of contents comprising:

searching page data to create a list of links in the given hyperdocument;

analyzing each link in conjunction with each other link in the list of links to identify link pairings;

assembling   link   pairings   in   order   to   form   clusters   of   links;

examining the links in the cluster of links for locality;

weeding out the links from the cluster of links which have properties that are not characteristic of intra-document links, to provide a resultant table of content set of identified candidate document pages; and,

grouping the content found in the resultant table of content set of candidate document pages by an automated system into a document representation stored in memory by the automated system; and,

printing, or viewing on a display by a user, the document representation.

2.   (Original)   The method of claim 1 wherein the step for analyzing each link further comprises determining a score for each link pairing.

3.   (Original)   The method of claim 2 wherein the scoring is determined by a proximity criteria.

4.   (Original)   The method of claim 2 wherein the scoring is determined by a similarity criteria.

5.   (Original)   The method of claim 2 wherein the scoring is determined by a regularity criteria.

6.   (Previously Presented)   A system identification methodology for assembling a document representation for subsequent viewing or printing of a given hyperlinked hyperdocument by gathering related hyperlinked page content comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

analyzing each link in conjunction with each other link to identify link pairings;

assembling link pairings in order to form clusters of links; and,

examining the links in the cluster of links for locality;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective table of content set of identified candidate document pages is assembled;

performing a document-level analysis that examines the collective table of content set of identified candidate document pages for grouping into one or more documents;

examining the collective table of content set of identified candidate document pages to weed out links from the collective table of content set which have properties that are not characteristic of intra-document links, to provide a resultant set of identified candidate document pages; and,

grouping the content found in the resultant set of candidate document pages by an automated system into a document representation stored in memory by the automated system; and,

printing, or viewing on a display by a user, the document representation.

7.   (Original) The method of claim 6 wherein the step for analyzing each link further comprises determining a score for each link pairing.

8.   (Original) The method of claim 7 wherein the scoring is determined by a proximity criteria.

9.   (Original) The method of claim 7 wherein the scoring is determined by a similarity criteria.

10. (Original) The method of claim 7 wherein the scoring is determined by a regularity criteria.

11. (Previously Presented) A system identification methodology for assembling a document representation for subsequent viewing or printing of a given hyperlinked hyperdocument by gathering related hyperlinked page content comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

searching page data to create a list of links in the hyperdocument;

analyzing each link in conjunction with each other link in the list of links to identify link pairings;

assembling link pairings in order to form clusters of links; and,

examining the links in the cluster of links for locality;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective table of content set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective table of content set of identified candidate document pages for grouping into one or more documents

examining the collective table of content set of identified candidate document pages to weed out links from the collective table of content set which have properties that are not characteristic of intra-document links, to provide a resultant set of identified candidate document pages; and,

grouping the content found in the resultant set of candidate document pages by an automated system into a document representation stored in memory by the automated system; and,

printing, or viewing on a display by a user, the document representation.

12. (Original) The method of claim 11 wherein the step for analyzing each link further comprises determining a score for each link pairing.

13.		(Original)	The method of claim 12 wherein the scoring is determined by a proximity criteria.

14.		(Original)	The method of claim 12 wherein the scoring is determined by a similarity criteria.

15.		(Original)	The method of claim 12 wherein the scoring is determined by a regularity criteria.

EVIDENCE APPENDIX

A copy of each of the following items of evidence relied on by the Appellant is attached:

DECLARATION UNDER 37 CFR §1.132 by Steven J. Harrington, Ph. D., filed 11/16/07.
The evidence was entered into the record by the Examiner in the 01/09/08 Office Action.

DECLARATION UNDER 37 CFR §1.132 by James M. Sweet, filed 11/16/07.
The evidence was entered into the record by the Examiner in the 01/09/08 Office Action.

RELATED PROCEEDINGS APPENDIX

NONE

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventor(s): James M. Sweet et al.

Application No.: 10/608,591

Filed: 6/27/2003

Examiner: Nathan Hillery

Art Unit: 2176

Confirmation No.: 8445

Title: **DETERMINATION OF TABLE OF CONTENT LINKS**
FOR A HYPERLINKED DOCUMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

DECLARATION UNDER 37 CFR §1.132

I, **James M. Sweet**, declare and state:

1. I am one of the inventors listed on the above-identified application. I reside at 394 Warren Ave, Rochester, NY, 14618.

2. I have a Bachelors and a Masters degree in Computer Engineering from Rochester Institute of Technology, Rochester, NY, USA, both received in 2003.

3. I have been employed by Xerox Corporation since 1998, where my current title is Member Research & Technology Staff II in Xerox Innovation Group. Since I joined Xerox, I have performed research and development in the area of image compression and image processing in software, and have generally specialized in software throughout both my college education and career at Xerox.

4. I have read and understand the contents of the U.S.P.T.O. Official Action of August 3, 2007, and am of the position that: a) in claims 1, 16, and 27, it would be clear to anyone skilled in the art that the "document representation" referred to in the aforementioned claims must implicitly refer to a document representation stored in computer memory; b) in claims 1, 16, and 27, it would be clear to anyone skilled in the art that the "subsequent viewing and printing" would be initiated by a user of a computer system; c) the claims being made in present application are distinctly different from and not obvious over Bharat et al. (U.S. Patent 6,112,203, hereinafter "Bharat") in view of Earl et al. (U.S. Patent 5,924,104, hereinafter "Earl"); and, d) the definition of "intra-document link" as meant in Earl, is entirely separate from the definition as meant in the present application, and that therefore Earl should not be considered in relation to the present application.

5. Claims 1, 16, and 27 each refer to "grouping the resultant set of candidate document pages into a document representation for subsequent viewing or printing of the given hyperlinked document." It would be understood by anyone skilled in the art that any meaningful representation of a hyperlinked document would necessarily have to be stored in memory, due to the inherent nature of hyperlinking. Merrian-Webster defines a "hyperlink" as "an electronic link providing direct access from one distinctively marked place in a hypertext or hypermedia document to another in the same or a different document." It would be fundamentally impossible for any document not stored in memory to contain a hyperlink, since by definition a hyperlink is an electronic link providing direct access to another document or place in the same document. This type of electronic direct access is impossible in any document not stored in memory. Therefore, because the "document representation" is said in the claim to represent a "hyperlinked document," it is implicit in the claim that the document representation must be stored in memory.

6. Furthermore, the "subsequent viewing or printing of the given hyperlinked document" would be reasonably understood by anyone skilled in the art to have been initiated by the user of a computer system. While it would be theoretically possible for a human to manually translate the coded representation of the hyperlinked document into a viewable form, this is impractical to the point of absurdity. It would be understood by anyone skilled in the art that "viewing or printing" a "hyperlinked document" would be an activity performed by a computer and initiated by a user of that computer.

7. It is my understanding that the Examiner believes the "intra-document links" referenced in several of the claims are obvious from Bharat, because both inventions take into account whether two nodes reside on the same server when determining to retain a link between those nodes. I believe this interpretation to be incorrect for a number of reasons. It should first be noted that the links being retained in Bharat have a very different meaning than the links being retained in the present invention. In Bharat, the intention is to improve the ability of a search engine to retrieve pages which are "important" in the context of the subject matter. Bharat has wisely observed that one's opinion of one's own importance is perhaps less accurate than others' opinions of one's importance, and has thence chosen to completely discard links that reside within the same server. Those skilled in the art will immediately recognize this technique, similar to the PageRank algorithm used by Google, as a means to prevent a common search engine exploit, wherein a website artificially increases its own perceived importance by linking to itself many times. In contrast, the links being retained in the present invention are meant to indicate a very specific and close relationship between hyperlinked pages, i.e. that they are part of the same continuous document. Bharat seeks to retain only those links which do not have a close relationship, while the present invention seeks to retain only those links which do have a specific type of close relationship.

-3-

8. Much discussion has been made of the use the phrase "weed out" in the present application, used on page 7, lines 9-22. It is my understanding that the Examiner does not believe that the language "weed out" implies removal of the links in question, asserting instead that "weed out" only implies identification, not removal. The full claim language phrase in question is "weed out links which have properties that are not characteristic of intra-document links." It should be noted that the following claim language makes clear that the intention of this step is to produce "a list of intra-document links." It defies the imagination to come up with any reasonable interpretation that, seeking to produce a "list of X," would first go to great pains to identify items which "have properties that are not characteristic of X," and yet then proceed to just carelessly toss those same items into the final "list of X" anyway, disregarding the previous identification.

9. I would also point to the present application, page 10, lines 23-27, where the applicants refer to the identification of hyperlinks that "are significantly different in a property that is typical of intra-document links," going on to again refer to the "link to a page on a different server" as being one of these criteria. In this part of the specification, the applicants used the phrase "filtered out," as an alternative to the "weed-out" language which will also further reinforce for those skilled in the art as to what the Applicants mean. Regardless, it would be clear to any skilled in the art that the Applicants are seeking to remove links that do not fit the characteristics of intra-document links.

10. Having established that Bharat uses the criteria of co-residence on the same web server as primarily an exclusionary criteria, while the present invention uses it as primarily an inclusionary criteria, one not skilled in the art may at this point be tempted to say that the links being retained in Bharat are merely the inverse of the links being retained in the second invention. This perception would be false. Recognizing that a close relationship may artificially inflate importance, Bharat seeks to eliminate *all* types of close relationships, and therefore, the inverse would be to include all types of close relationships indiscriminately. The present invention seeks to identify a *specific type* of

close relationship (the relationship of being part of the same document) and therefore the inverse would include a number of links which also identified a close relationship of a different type, i.e. a different set of links than that which is taught by Bharat.

11. One skilled in the art would further recognize that whether or not two given hypertext pages reside on the same web server is merely a piece of data, and not a teaching in and of itself. To suggest that Bharat makes obvious the present invention because both sample this piece of data is analogous to suggesting that a design for a central air conditioning system makes obvious a design for a nuclear reactor, because both inventions require the measurement of temperature in order to maintain proper operation.

12. In short, even if the Applicants' definition of "weed out," is or is not accepted, it is nevertheless irrelevant to whether Bharat makes the present invention obvious. Making a determination based on whether two pages reside on the same web server, is not original to the teachings of Bharat, nor the present Applicants, as it is a piece of data that is used everyday in a number of different applications, such as Domain Name Server lookups, corporate firewalls, page caching mechanisms, web browser pop-up blockers, etc. It will be clear to those skilled in the art that many unique inventions will at times sample some of the same data. It is the different ways in which this and other data points are interpreted that is central to both Bharat and the present invention and which is also where the distinction between them is to be found.

13. It is unfortunate that both the present Applicants and Earl have chosen to use the term "intra-document link." It will be clear to anyone skilled in the art that this is merely a semantic coincidence, stemming from a different meaning of the word "document."

14. The present Applicants define an intra-document link as "links within a given web page that may link to *other* pages within the same document" (emphasis added).

This makes it clear that a "document" as intended by the current applicants may refer to a group of more than one web page, since it would be impossible to link to other pages within a document if a document only ever comprised a single page.

15. Earl does not ever specifically define "document" or "intra-document link," however, one skilled in the art can clearly discern what is meant by "document" in Earl's teachings. In Earl, Column 3, Line 44 through Column 4, Line 15, five examples are given of links which may be "intra-document." Earl observes that only the first two can reliably be said to be "intra-document," while all five could be "intra-document." One skilled in the art will readily observe the point Earl is making: In all five examples, the link appears to be a link to the same web page as the one in which the link appears. However, due to the details of how link destinations are interpreted by web servers, only the first two examples are certain to link to the same web page as the one in which the link appears. It is abundantly clear from this that what Earl means by "intra-document" is a link whose destination is the same as its source, i.e. a circular link to the same page (hereinafter referred to as "circular link").

16. One skilled in the art will recognize that Claim 1 of the present application both implicitly and explicitly excludes circular links from any form of consideration. Claim 1 teaches of "a recursive application of page-level link analysis to the linked candidate document page and any *further* nested candidate document pages thereby identified" (emphasis added). The word "further" explicitly excludes circular links from consideration, because they would not increase the number of candidate document pages identified. In addition, the "recursive application" will be recognized by one skilled in the art to implicitly exclude circular links, since the inclusion of circular links in such an analysis would result in infinite recursion.

I, the undersigned, further declare that all statements made herein are of my own knowledge are true and that all statements made on information and beliefs are believed to be true; and further, that these statements were made with the knowledge that willful

false statements and the like so made are punishable by fine or imprisonment, or both under Section 1001 of Title 18 of the United States Code, and further such willful statements may jeopardize the validity of the application or any patent issuing thereon.

Respectfully submitted,

James M. Sweet

11/13/07

Date

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventor(s): James M. Sweet et al.

Application No.: 10/608,591

Filed: 6/27/2003

Examiner: Nathan Hillery

Art Unit: 2176

Confirmation No.: 8445

Title: **DETERMINATION OF TABLE OF CONTENT LINKS FOR A HYPERLINKED DOCUMENT**

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

DECLARATION UNDER 37 CFR §1.132

I, **Steven J. Harrington, Ph. D.**, do hereby declare and state:

1. I am one of the inventors listed on the above-identified application. I reside at 251 Burnett Road, Webster, New York 14580.

2. I have Bachelors degrees in mathematics and physics from Oregon State University in Corvallis Oregon, received in the year 1968, and Masters degrees in physics and computer science from the University of Washington, in Seattle Washington, received in year 1969 and year 1976, respectively. I then also received a Ph.D. in physics from the University Of Washington, in Seattle Washington in year 1976.

3. I have been employed by Xerox Corporation as a scientist and inventor for 26 years, where my current title is Research Fellow, in the Xerox Research Center Webster of the

Xerox Innovation Group. Since I joined Xerox, I have performed research and development in the area of document engineering and digital imaging technologies and have been granted over 120 patents in those technology areas.

4. I have read and understand the material provided with regard to patent application No. 10/608,587 by Sweet et al. including the application itself, the amended claims, patents 6,112,203 to Bharat, 5,924,104 to Earl and 6,877,002 to Prince, as well as the remarks and arguments to the patent examiner the contents of the U.S.P.T.O. Official Action of August 3, 2007, and am of the position that a) claims 1-6, 10-13, 16-20, 25-31, 36, and 37 are not obvious over Bharat et al. (U.S. Patent 6,112,203) in view of Earl, (U.S. Patent 5,924,104), and that claims 7-9, 14, 15, 21-24 and 32-35, are not obvious over Bharat et al., in view of Earl, in view of Prince, (6,877,002).

5. The teachings of Earl deal with intra-page links even though they are referred to as intra-document links. As such they would be removed from consideration by the method as taught and claimed in the present Application. The Earl patent does nothing to provide enlightenment towards the problem addressed by the present Application. Nor does the Prince patent provide teachings that address the shortfalls of Bharat and Earl. Bharat is attempting to assemble a set of distinct documents relevant to a search topic, while the present Application is teaching the identification of links to components of a single document. As such, the Bharat method quickly discards the very links that the present Application is seeking to identify. Furthermore, since Bharat's method rejects many documents and web pages besides those belonging to a document, one could not simply collect the referenced pages that Bharat rejects. Some additional points are that Bharat is working from a set of documents identified by a search engine and therefore has no specified preferred document for use in the identification of its hyperlinked components. Furthermore, since Bharat is working on a set of document identified by a search engine, there is not guarantee of connectedness of the resulting n-graph. None of the teachings of Bharat do anything to detect or enforce connectedness, while for the problem of document boundary identification addressed by the present Application, connectedness of the graph is central, and the process taught of following links chained from an original

source page, guarantees it. While both Bharat and the present Application examine the same space of hyperlinked web pages, and both seek to classify web pages identified in that space, their goals are quite different, leading to different search, analysis, and classification techniques. The teachings of Bharat cannot be used to solve the problem addressed by the present Application, and do not anticipate the teachings of the present Application.
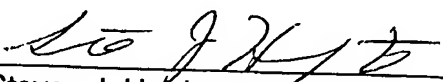
6. With regards to the amendment of the claims as to a document representation "stored in memory". We the Applicants are dealing with exactly the same document representations as are Bharat and Earl, namely web pages provided by a web server. It will be understood by one skilled in the art that a web server is a computing device whose purpose is to deliver web pages to web browser's over a communications channel such as the internet. It will be understood by one skilled in the art that a web server consists of a processor and memory. It will be understood by one skilled in the art that the memory of the web server contains the web page document in a representation from which the processor can copy, transform or otherwise generate the form understood by the web browser that is requesting the web page. The present application claims "an automated identification methodology for assembling document related hyperlinked pages". It further describes and claims "an automated document boundary detection system". As one skilled in the art, it is obvious to me that automation of the method entails a computing system capable of retrieving and analyzing the electronic representations of the web pages in accordance with the methods taught by the application specification. Such a computing system would have a processor able to conduct the analysis and memory to hold the document representation and grouping results. I therefore believe that the amendments to the claims indicating that the resultant set of document pages are grouped into a document representation stored in memory, is supported by the specification in the light of the digital electronic nature of the documents and the indication of automated processing in the claims, and would be so understood as such by those skilled in the art.

7. As to the term "weed out", this will be understood by one skilled in the art as meaning to discriminate *and discard*. Gardening terms such as "weed out" or "prune" are not uncommon in computer science and are used with usual English understanding of eliminating from further consideration. In computing, this often includes removal of the weeded object from memory, but whether or not this is the case, the weeded object will no longer be included among the objects being processed. This meaning can be seen in the specification which states "...links 240 are passed to module 250 for a final examination to weed out links which have properties that are not characteristic of typical intra-document links..." and then in the next sentence state "The final result is then a list of intra-document links 120 for the candidate page 270". In other words, the links that are not intra-document links are not only distinguished, but also discarded by the weeding process, leaving just the intra-document links.

8. I, the undersigned, further declare that all statements made herein are of my own knowledge are true and that all statements made on information and beliefs are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both under Section 1001 of Title 18 of the United States Code, and further such willful statements may jeopardize the validity of the application or any patent issuing thereon.

Signed:

Respectfully submitted,

Steven J. Harrington, Ph.D.

November 6, 2007
Date

-4-